Divergence Note: Architecture-Dependent Effects of Toroidal Attention Constraints

Sylvain Cormier
Paraxiom Research
sylvain@paraxiom.org

January 16, 2026

Abstract

We report a critical finding from experiments validating toroidal (Tonnetz) attention constraints for hallucination reduction: **the same constraint that reduces hallucination in Phi-2 by 50% increases hallucination in TinyLlama by 180%.** This divergence invalidates universal applicability claims and opens fundamental research questions about architecture-topology compatibility. This note serves as an immediate disclosure of negative results that materially affect interpretation of prior positive findings.

Related paper: "Topological Constraints for Coherent Language Models" (DOI: 10.5281/zenodo.18187835)

Status: Preliminary findings requiring immediate disclosure

1 Summary

We tested toroidal attention constraints on two language models without RLHF alignment to isolate architectural effects. The identical constraint produces opposite effects:

- **Phi-2 (2.78B):** 50% hallucination *reduction*
- **TinyLlama (1.1B):** 180% hallucination *increase*

This is not noise—the TinyLlama result is consistent across 50 samples.

2 Method

We inject a log-space bias into attention scores before softmax, derived from toroidal (Tonnetz) distance on a 12×12 torus:

```
# Tonnetz topology: 12-tone musical lattice on 2D torus
def toroidal_distance(i, j, grid_size=12):
    xi, yi = i % grid_size, (i // grid_size) % grid_size
    xj, yj = j % grid_size, (j // grid_size) % grid_size
    dx = min(abs(xi - xj), grid_size - abs(xi - xj))
    dy = min(abs(yi - yj), grid_size - abs(yi - yj))
    return dx + dy

# Attention mask: full weight within radius, decay outside
mask[i,j] = 1.0 if distance <= radius else exp(-alpha * distance)
```

```
# Injection: log-space bias added before softmax
topo_bias = log(mask + 1e-10)
attention_scores = attention_scores + topo_bias + causal_mask
```

Parameters: radius = 2.0, $\alpha = 1.0$, grid_size = 12

3 Results

3.1 Phi-2 (2.78B parameters, NO RLHF) — Positive Result

Condition	TruthfulQA	HaluEval	Spectral CV
Baseline	0%	40%	7.39
Toroidal	0%	**20%**	7.68

Table 1: Phi-2 results ($n = 10$, reproduced 3×). **Effect: 50% hallucination reduction.**

3.2 TinyLlama (1.1B parameters, NO RLHF) — Negative Result

Condition	TruthfulQA	HaluEval	Spectral CV
Baseline	0%	10%	7.55
Toroidal	2%	**28%**	7.60

Table 2: TinyLlama results ($n = 50$). **Effect: 180% hallucination INCREASE.**

4 Critical Warning

Same constraint. Same parameters. Opposite sign.
The toroidal topology that improves Phi-2 actively harms TinyLlama. This is not statistical noise—the effect is large and consistent across 50 samples.

5 Interpretation

Possible explanations under investigation:

1. **Model capacity:** Larger models (2.78B) may absorb constraints better than smaller ones (1.1B)
2. **Training data:** Phi-2's "textbook-quality" data may align with structured geometric constraints; TinyLlama's web-scraped data may not
3. **Topology mismatch:** The 12-tone Tonnetz periodicity may suit certain attention patterns but destructively interfere with others
4. **Attention head distribution:** Different architectures may require different topologies

6 Implications

1. **No universal fix:** Geometric constraints cannot be applied blindly across architectures
2. **Architecture-specific topologies:** Research must identify which topology fits which model
3. **Validation required:** Any topological intervention must be validated per-architecture before deployment

7 Reproducibility

```
# Environment
Python 3.11, PyTorch 2.1, transformers 4.36
Random seed: 42

# Phi-2 (positive result)
python phi2_definitive_proof.py --model phi-2 --mode quick

# TinyLlama (negative result)
python phi2_definitive_proof.py --model tinyllama --mode full --samples
    50
```

All code and results: `https://github.com/Paraxiom/topological-coherence`

8 Next Steps

1. Hyperparameter sweep: test radius $\in \{1, 2, 4, 6\}$, $\alpha \in \{0.3, 0.5, 1.0, 2.0\}$
2. Alternative topologies: linear distance, different grid sizes
3. Layer-specific constraints: wrap only early/late layers
4. Larger sample validation on GPU infrastructure

Citation

If referencing this finding:

> Cormier, S. (2026). Divergence Note: Architecture-Dependent Effects of Toroidal Attention Constraints. *Paraxiom Research*. January 16, 2026. Supplement to DOI: 10.5281/zenodo.18187835

Contact: @ParaxiomAPI | `research@paraxiom.io`